Homework 3: MATLAB Portion

This section contains the code and results for problem 6. The results section address each subpart of the question.

Code

The following is the code that was used to solve problem 6.

```
%% ECEM146
% Author: Thomas Kost
% UID: 504989794
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% HW3:

%accuracy reference

x_train = dlmread('dataTraining_X.csv');
y_train = dlmread('dataTraining_Y.csv');

x_test = dlmread('dataTesting_X.csv');
y_test = dlmread('dataTesting_Y.csv');

%calculate accuracy

train_sum = sum(y_train);
test_sum = sum(y_test);

majority_train = 1;
majority_test = 1;
if (train_sum < length(y_train) -train_sum)
    %0 dominant
    train_sum = length(y_train)-train_sum;
    majority_train= 0;
end
if(test_sum < length(y_test) -test_sum)
    %0 dominant
    test_sum = length(y_test)-test_sum;
    majority_test = 0;
end

train_accuracy = train_sum/length(y_train);
test_accuracy = test_sum/length(y_test);


%create decision tree
tree = fitctree(x_train, y_train, 'SplitCriterion','deviance');
y_predict_training = tree.predict(x_train);
y_predict_test = tree.predict(x_test);

accuracy_train(1:length(y_train)) = y_train==y_predict_training;
```

```
accuracy_test(1:length(y_test)) = y_test==y_predict_test;

Train_result = sum(accuracy_train)/length(accuracy_train);
Test_result = sum(accuracy_test)/length(accuracy_test);
```

Results

Part A

We were able to calculate a baseline accuracy for each data set. This value was the proportion of the labels of the set that belong to the majority class. In calculating this we found a baseline accuracy for the training set of 0.5944 and a baseline accuracy for the testing set of 0.6949.

Part B

We then used the builtin fitctree function to train a decision tree on our training data. We then used this model to predict the labels for both the training and testing data. This resulted in the tree achieving a 0.9099 training accuracy and a 0.8192 testing accuracy. These values suggest that our model is doing a good job of classifying points (as they are quite higher than the baseline accuracy), but it also suggests that we have not overfit the model either. This is because both the training data and the testing data produced fairly high accuracy results, so our model extends to the unseen points.